Predicting U.S. Domestic Air Travel Delays

Patrick Sollars, Aaron Newman, and Cecilia Chen

Introduction

No one likes flight delays. Unfortunately, they continue to happen, and while they are rare, they happen more frequently than we would prefer. The goal of this project is to better understand the leading justifications of delays in United States domestic air travel and whether we can effectively predict the likelihood of a delay in the future.

The Datasets

Our primary dataset for this effort will be Airline On-Time Performance Data managed by the U.S. Department of Transportation's Bureau of Transportation Statistics (BTS). On a monthly basis, participating U.S. airlines report the status of each one of their flights to BTS, which includes the actual times of departure and arrival, the scheduled times of departure and arrival, and a host of additional information about the flight. This information includes the flight number, origin and destination, and the registration number (also called tail number or N-number) of the aircraft. If the flight did experience a delay, information is provided about the cause such as carrier, weather, National Air System (NAS), security, or late inbound aircraft. More information about this dataset can be found here: BTS Airline On-Time Performance Data.

We are also using the Federal Aviation Administration's (FAA) Releasable Aircraft Database to match registration numbers from the BTS data with information about the specific aircraft, to include airframe type, engine type, and year of manufacture. Additional details about this dataset can be found here: FAA - Aircraft Registration Database Download.

One challenge of our chosen datasets is the sheer number of records. We will have to contend with handling several Gigabytes of information in order to perform our analysis. While we intend to leverage the Great Lakes computing cluster, it will also be important to consider the computational cost of our modeling techniques.

We also need to consider the time range of records for this analysis. We are hesitant to incorporate data from 2020-2022 which would include the pronounced effects of the pandemic on airline travel. Our goal is to analyze air travel under "normal" circumstances as much as possible. We will likely use records from a 5 year period of 2015-2019.

Prior Research

One of our team members participated in a Milestone I project using similar data from BTS to study the impact of COVID-19 infection rates on air travel passenger volumes. The project was able to use Vector Autoregression (VAR) to make reasonable predictions, but the monthly data used was not sufficiently granular to use more advanced time-series prediction techniques. This

Milestone II project will take advantage of some of the knowledge gained, but is using a more granular dataset to answer different questions. More information about the Milestone I research can be found here: https://github.com/newmanar/SIADS593.

We have discovered that there are quite a few papers, articles, and Kaggle projects using the BTS On-Time Performance Data. One such Kaggle tutorial was published in 2017 by Daniel Fabien [1]. We were disappointed to discover a paper that is still available online that appears to directly use information from the Kaggle tutorial without citation. [2] Another instructive article was published as a Medium post in 2020. [3]

We intend for our Milestone II research to be distinct from prior published research in the following ways:

- The effective use of unsupervised techniques to assist in feature engineering
- Broader use of data not limited to a specific geography
- The use of aircraft type, engine type, and aircraft age as potential features
- The use of time-series prediction techniques as part of our analysis

Unsupervised Learning

Examining the "BTS Airline On-Time Performance" dataset, it seems the primary causes of flight delays are attributed to carriers and late aircraft, which are related to airline operations and airport logistics. Furthermore, shared flights tend to experience more extended delays in both departure and arrival times when compared to their non-shared flights.

Utilizing clustering techniques can enhance our comprehension of the flight dataset by revealing underlying groups or patterns. Given that the dataset exhibits a non-nested structure, K-means (medians) will be initially considered for discovering partitions. We will also explore alternative unsupervised clustering algorithms such as DBSCAN.

Both marketing and reporting flight data sets include more than 100 features. To prepare the data for constructing predictive models, we intend to employ dimensionality reduction using Principal Component Analysis (PCA) for linear feature relationships. It may also be worth exploring techniques such as Kernel PCA if we begin to suspect there is a non-linear feature relationship. Tools like T-SNE will be helpful to visualize our feature landscape before analysis. We will assess the effectiveness of the mentioned techniques by evaluating their impact on the performance of our predictive model using cross-validation. Subsequently, we will choose the most suitable approach for our model.

Supervised Learning

Informed by our unsupervised learning work to assist with feature selection and engineering, we will train models using several techniques including, but not limited to Naive Bayes, Random Forest, or Support Vector Machines. We will compare performance between these classification models using a range of evaluation metrics such as accuracy, precision, and recall. We will also look to aggregate the available data at the daily or weekly level and apply techniques such as Seasonal Autoregressive Integrated Moving Average (SARIMA) or SARIMA with Exogenous Factors (SARIMAX) to be able to predict aggregate delays as a function of time. Metrics like root mean square error will help evaluate performance on these time series based models. Due to the inherent uncertainty in real-world data and the complexity of flight delays, model performance may vary. Therefore, we will also assess the robustness of our models through techniques such as cross-validation and sensitivity analysis to account for this uncertainty.

Team Contributions

Note: All team members will participate in all phases of the analysis. Named team members will serve as leads/coordinators of each portion of the research.

- Initial proposal - Full Team
- FAA Exploratory Data Analysis - Patrick
- BTS Exploratory Data Analysis - Aaron
- Selecting Unsupervised Learning Techniques - Cecilia
- Selecting Supervised Learning Techniques - Aaron
- Identifying and Selecting Computing Environment(s) - Patrick
- Data Preparation for Analysis - Patrick
- Perform clustering and synthesize results - Cecilia
- Perform non-temporal supervised learning work and synthesize results - Patrick
- Perform time-series predictive model work and synthesize results - Aaron
- Initial draft of final report - Aaron
- Final report completion - Full Team

Timeline

- Aug 29 - Sept 4 - Project alignment, Draft proposal
- Sept 5 - Sept 11 - Proposal revisions, Set up Great Lakes computing environment
- Sept 12 - Sept 18 - Exploratory data analysis, Data preparation
- Sept 19 - Sept 25 - Model exploration, evaluation and selection
- Sept 26 - Oct 2 - Unsupervised and supervised modeling, results analysis (Stand-up 1)
- Oct 3 - Oct 9 - Time series analysis, Predictive modeling
- Oct 10 - Oct 16 - Drafting report, Building data visualizations, (Stand-up 2)
- Oct 17 - Oct 23 - Polish and submit final report

References

[1] Fabien, Daniel, Predicting flight delays [Tutorial], Kaggle, 2017. https://www.kaggle.com/code/fabiendaniel/predicting-flight-delays-tutorial. (Accessed 1 September 2023.)

[2] Prabakaran N. and Rajendran Kannadasan, Airline Delay Predictions using Supervised Machine Learning, International Journal of Pure and Applied Mathematics, 119(7), February

2018. https://www.researchgate.net/publication/325034541_Airline_Delay_Predictions_using_Supervised_Machine_Learning. (Accessed 1 September 2023.)

[3] Herbas, Javier. Using Machine Learning to Predict Flight Delays, Analytics Vidhya, 17 October 2020. https://medium.com/analytics-vidhya/using-machine-learning-to-predict-flight-delays-e8a50b0bb64c. (Accessed 1 September 2023)